DOI:
10.1002/asi.23767

Document Version
Peer reviewed version

Insight workflow: Systematically combining human and computational methods to explore textual data

Alastair J. Gill, (corresponding author)

Department of Digital Humanities, King's College London,

26-29 Drury Lane, London WC2B 5RL

Tel. +44 (0)20 7848 1979, Fax. N/A

alastair.gill@kcl.ac.uk

Saba Hinrichs,

The Policy Institute at King's, King's College London, Virginia Woolf Building,

22 Kingsway, London, WC2B 6LE

Tel. +44 (0)20 7848 7982, Fax. N/A

saba.hinrichs@kcl.ac.uk

Tobias Blanke

Department of Digital Humanities, King's College London,

26-29 Drury Lane, London WC2B 5RL

Tel. +44 (0)20 7848 1975, Fax. N/A

tobias.blanke@kcl.ac.uk

Jonathan Grant

The Policy Institute at King's, King's College London, Virginia Woolf Building,

22 Kingsway, London, WC2B 6LE

Tel. + 44 (0) 20 7848 1742, Fax. N/A

jonathan.grant@kcl.ac.uk

Mark Hedges

Department of Digital Humanities, King's College London,

26-29 Drury Lane, London WC2B 5RL

Tel. +44 (0)20 7848 1970, Fax. N/A

mark.hedges@kcl.ac.uk

Simon Tanner

Department of Digital Humanities, King's College London,

26-29 Drury Lane, London WC2B 5RL

Tel. +44 (0)20 7848 1678, Fax. N/A

simon.tanner@kcl.ac.uk

Abstract

Analysing large quantities of real world textual data has the potential to provide new insights for researchers. However, such data present challenges for both human and computational methods, requiring a diverse range of specialist skills, often shared across a number of individuals. In this paper, we use the analysis of a real world data set as our case study, and use this exploration as a demonstration of our 'insight workflow', which we present for use and adaptation by other researchers. The data we use are impact case study documents collected as part of the UK Research Excellence Framework (REF), consisting of 6,679 documents and 6.25 million words; the analysis was commissioned by the Higher Education Funding Council for England (published as report HEFCE 2015). In our exploration and analysis we used a variety of techniques, ranging from keyword in context and frequency information to more sophisticated methods (topic modelling), with these automated techniques providing an empirical point of entry for in depth and intensive human analysis. We present the 60 topics to demonstrate the output of our methods, and illustrate how the variety of analysis techniques can be combined to provide insights. We note potential limitations and propose future work.

Introduction

The greater the available data, the greater the potential insights that can be gained from it; however, even in the exploratory stages of understanding what data we have available to us, we need to carefully choose analysis techniques in order to accurately understand our data. Considering textual data specifically, this paper provides a step-by-step framework for other researchers interested in applying corpus exploration and natural language processing techniques – including in particular topic modelling - to large amounts of textual information: to aid and facilitate exploration and both qualitative and quantitative analysis of this data; and to best target in-depth human coding. We demonstrate how this combination of approaches to text analysis provide us with a greater and more rounded understanding of large sets of documents, and we provide detailed steps to facilitate the application or adaptation of our 'insight workflow' by others. Our dataset is the 'impact case study' documents submitted as part of the UK Research Excellence Framework (REF) in 2014, which was administered by the Higher Education Funding Council for England (HEFCE). Our analysis was commissioned by HEFCE, with the final report describing the results of our analysis published in HEFCE (2015). In the current paper, we describe in more detail the methods that we adopt for such exploratory analyses, and use the HEFCE analysis as a worked example.

Although analysis methods are often distinguished by groupings such as qualitative/quantitative, hypothesis driven/data driven, or computational/manual, we believe such divisions to be unhelpful and in our insight workflow we propose a combined approach, capturing the strengths of each (cf. Mehl and Gill, 2010; Oberlander and Gill, 2006). One useful way of describing text analysis methods is in terms of 'effort', and in particular where the majority of the human input

is required in a particular analysis technique (Quinn et al. 2010): For example, human input may be required to construct particular categories or codes (e.g. 'dictionaries', Vasalou et al., 2011; Pennebaker et al., 2003) and thus have high pre-analysis costs, which can then be automatically applied to large data sets with little human involvement (low analysis and post-analysis costs); in contrast, corpus linguistics or natural language processing methods which extract raw word frequency information or instances of particular words or grammatical categories, such as nouns, can perform this analysis at very little human cost, but it is the subsequent post-analysis required to interpret these results which require high levels of human involvement (unsurprisingly, manual reading or coding has high labour costs at all stages of analysis). Although computational approaches such as text mining have been heralded enthusiastically among humanities and social science scholars (e.g., Kirschenbaum, 2007; Borgman, 2009, Graham et al. 2015), a major issue remains that many analysis tools and techniques exist, but for the most part the majority of skill is required in knowing when to apply them and how to combine them with data to harness the strengths of each approach. This challenge is the focus of the current paper.

One of the key analytic tools which we apply to our data is topic modelling (Blei, 2012), an approach which has gained significant attention in the arts, humanities and social sciences in recent years. It is a data-driven technique, but to some extent it sits between the extremes in terms of how much human input is required at the start of analysis to match the topic model to the data versus interpreting the output following analysis ([insert footnote 1 here]). Topic modelling has been applied to a wide range of research questions, such as politics, history, and scientific publishing and practices, and has become popular in historical research to go through large-scale archives (Klein et al. 2013; Blei, 2012). For historians, automated topics have even

been described as the 'hands-on adventure' in Big Data (Graham et al. 2015). Topic models have also been used to track social media dynamics and assess policy decisions in order to overcome subjective assumptions (Brauer and Dymitrow, 2013). Indeed, one of the most relevant applications to the data set analysed in this paper has been to group NIH funded research projects by their topic rather than by formal categories, for use in a searchable database (Talley et al. 2011). One of the great strengths of topic modelling is in its ability to group words together into (generally) interpretable categories based on their patterns of occurrence within a set of documents (relevant to techniques from information retrieval, and proposals of human language ability; Landauer and Dumais, 1997). We do, however, note that this is not always the case, and that the 'cohesion' between terms forming a topic is one indicator used to identify a topic model which is well-fitted to the data (we discuss this and other evaluation methods in our 'insight workflow' description of topic modeling). Furthermore, since it is based on unsupervised learning, it can be quickly applied to an unknown body of documents, making it potentially well-suited to data containing spelling or orthographic variations (e.g., online text), or instances where word meanings may differ to evolve over time (e.g., Rule et al. 2015).

Although topic modelling is a powerful technique, care needs to be taken in its application and in the interpretation of its results, for example, there has been criticism about applicability of insights gained (Sula, 2013) and the challenge of developing a reliable non *ad hoc* evaluation framework (Chang et al. 2009). One proposal has been to develop interactive topic modelling software, which proved beneficial to political scientists and this ability to understand large collections of data (Hu et al. 2014). In this paper, we begin to address such limitations of topic modelling: we do so by proposing a framework which uses standard or existing software and

approaches to enhance the understanding of a large unknown document collection. We demonstrate this approach using the impact case studies collected as part of REF 2014 in order to understand patterns of wider societal impact resulting from academic research. The rest of the paper is structured as follows: We provide greater detail on the data which we analyse, and also the automated text analysis approaches incorporated into our analysis; specifically, we will present a more general analytic pipeline that others will be able to adapt and apply to their own data in future. Then follows the results and evaluation of our analysis in the context of our findings; here we briefly describe results from the previously published study (HEFCE 2015) in order to illustrate and evaluate our approaches. Finally, we provide conclusions, and briefly note limitations and potential applications of this work.

Methods

In this section, we provide more information about the dataset we used for the analysis of the impact of UK higher educational institutions (HEIs) and note in particular why it is suited to our analytic approach. We then provide background information about the text analysis approaches we are adopting. Finally, we present our 'insight workflow' method to extract meaning from the case studies, and which we document as a step by step process for the benefit of other researchers.

Data

Through a contract with the Policy Institute at King's College, HEFCE provided us with advanced secure server access to the impact case studies of the REF 2014 (now available via: http://impact.ref.ac.uk/CaseStudies/; see http://www.hefce.ac.uk/rsrch/REFimpact/ for a greater

overview). Submissions to REF2014 (of which the impact case studies were a constituent part), were submitted to one of 36 disciplinary 'Units of Assessment', which in turn were grouped into four large categories or assessment 'Panels', namely: Panel A, Life sciences; Panel B, Engineering and physical sciences; Panel C, Social sciences; Panel D, Arts and humanities (see Table 10 in the Appendix for further information on their relationship to the 36 Units of Assessment). Examples of the kinds of impact described could include: the development of a super-repellent surface, created by plasmachemical techniques and invented by UK researchers, is used in millions of products worldwide, including mobile phones and hearing aids; Paralympic athletes' performance was improved by investigating wheelchair propulsion and optimizing configurations for competitive sport; research showing the importance of same-day diagnostic tests for tuberculosis led to improvements in access to care and reductions in costs incurred by patients in Malawi, Nigeria, Yemen, Ethiopia, Nepal and elsewhere; or editorial and biographic analysis of the work of Virginia Woolf directly fed into the composition of Vanessa and Virginia (2008), a novel by Susan Sellars about Woolf's relationship with her sister, Vanessa Bell (samples of these texts which were analysed in the current study can be found in Table 11).

Each impact case study had five sections: summary of impact; a description of the underpinning research; references to that research; details of the impact; and sources to corroborate the impact. We were given 6,679 non-redacted impact case studies that were submitted to the 2014 Research Excellence Framework (REF). Each case study aims to showcase how research undertaken in UK universities has benefited society beyond academia – whether in the UK or globally. The case studies outline changes and benefits to the economy, society, culture, public policy and services, health, the environment and quality of life. The documents were on average 2,142 words (ranging from 1,316 to 3,260 per document, and giving 14.31 million words in total), with

the current analysis limited to their fourth section (‘details of the impact’; an average 939.4 words per document, ranging from 147 to 2,152 words; 6.25 million words in total).

Given the scale of this very varied dataset, we developed our ‘insight workflow’, in order to not be restricted by existing frameworks and taxonomies of impact, which we identified as too conceptual and often limited in scope to a particular discipline. Developing many, individual impact taxonomies by hand was deemed too labour intensive given the time constraints of this project. Human coding of the documents alone would also have been prohibited by the scale of the data set and so topic modelling is the principal technique which we use to do this. To our knowledge, the study reported in HEFCE (2015) is the first of its kind in the domain of impact that uses an empirical approach based on such a breadth of material.

The original texts from the impact case studies were supplied in PDF format, which had to be cleaned and processed for subsequent analysis: in this case, we required plain text as input for our analysis, which was extracted from each document using the ‘pdftotext’ UNIX command line utility, with the ‘layout’ command option specified to maintain as much of the text structure as possible (in order to better extract document sections). Features of the documents removed (cleaned up) to improve analysis included section titles (e.g., ‘details of the impact’) and page numbers. Finally, in the case of the topic modelling analysis (TM1 [add footnote 2 here]), we note additional text processing steps: texts were lowercased, non-ASCII characters and punctuation were removed, and indicators of redacted text were consistently replaced with ‘xxxx’; for filtering and normalizing features, we removed general stop words (i.e. frequent, often grammatical words such as ‘the’, ‘a’, or ‘and’, which do not add meaning to the text; using

the list of 422 function words from McDonald, 2000) and also a small number of specific stop words identified as being very frequent in the case studies (impact, new, page, case, study, date, ref, research). The final preparation step for our topic modelling was 'stemming' the text, i.e. different forms of the same word (e.g., *runs, running, ran*) were reduced to a consistent or most basic form (e.g., in this case *run*) using the Porter stemming algorithm (http://tartarus.org/martin/PorterStemmer/), part of the Snowball package.

There is on-going debate considering the usefulness of stemming for topic modelling. In the preparation phase, we explored different stemming algorithms in preparation of the main experiment as well as not using stemming. One consideration, for example, is the aggressiveness of stemming and its impact on the further text analysis. The Porter stemmer is one of the best-known and packaged in NLP applications such as NLTK (nltk.org), while the Lancaster stemmer, also part of NLTK, is considered to be more aggressive, whereas there are other less aggressive stemmers (e.g., EnglishMinimalStemmer or lemmatisation algorithms). While most topic modelling examples we have seen in the literature use stemming in order to reduce our sizeable data set (e.g., Harvey et al. 2013; Wang et al. 2013; Jacobi et al. 2016), it is not always recommended, as word stems can be associated with the wrong topics. However, in our case we found the lexical diversity of our corpus (mean type-token ratio is 0.74 calculated over 200 word sections of text with stop words removed) to be big enough to justify stemming. In a corpus that is covering topics in distinct research disciplines the danger that stems are linked by expert reviewers to the wrong topic is limited and outweighed by the potential advantages of stemming such as the reduction of the data (we discuss this in more detail in the Results and Discussion section, below). To balance such trade-offs, is in fact one of the advantages of our 'insight

workflow' approach, where we integrate automated and manual text analysis and can be generous with the words linked to each topic. .

Computational Techniques used in the Analysis

In this paper, we focus on our analysis of the 'details of the impact' section, which consisted of free text. To analyse such a relatively large amount of data (at least, in human terms), we adopted three broad approaches to our analysis using text mining: topic modelling, keyword analysis and named entity extraction. Topic models aim to uncover hidden thematic structures or 'topics' that occur in a collection of documents utilising unsupervised machine-learning techniques (Blei, 2012). A topic consists of a cluster of words or phrases that show similar patterns of occurrence; documents may relate to more than one topic, and topic modelling calculates a weight with which each topic relates to a particular document. In contrast to topic modelling, keyword analysis – that is, searching for and/or counting occurrences of specific key words of interest - allowed us to look for specific instances of impact across a large number of texts; 'keyword in context' (KWIC) analysis enabled the viewing of list of keywords with n-characters/words of context displayed either side, which gives additional information to aid disambiguation of word sense or intended meaning of keywords. Named entity extraction was used to identify proper nouns such as countries, cities and institutions, with this generally working by matching items in the case studies against third party information. We typically used keywords to select a set of case studies that we read and analysed using qualitative approaches, such as close reading, or to gain a very general sense of the data. Keyword analysis, KWIC and information extraction for names and concepts were all implemented using freely available software, such as Python and NLTK, along with common UNIX command line tools, such as 'grep' and 'sort' (alternative

options for those not wishing to develop software include AntConc [Anthony, 2016] or OpenCalais [http://www.opencalais.com]). To compare our data against the British National Corpus we used Wmatrix (Rayson, 2008).

We used Latent Dirichlet allocation (LDA) for topic modelling (Blei 2012). As a generative technique, LDA starts with a model that is then used to describe the data by adjusting the parameters to fit the model. The assumption is that the whole corpus of documents contains k number of topics (specified by the user), and that each document talks about these k topics (to a greater or lesser extent). Therefore, each word in a document depends on both the topics selected for that document as well as the word distribution within each of these topics. This intuition is operationalized as a Bayesian Network that models this document generation process. We provide the following description of LDA to provide the reader with a better understanding of this much more opaque analysis stage. For further information on LDA, please consult Steyvers and Griffiths (2007), and for more detail of the steps involved in its practical application, see Mimno et al. (2014).

Using the notation of Steyvers and Griffiths (2007), LDA can be described as follows: $P(z)$ is the distribution of T number of topics over a particular document; $P(w|z)$ is the probability distribution of words w given topic z; $P(z_i = j)$ is the probability that topic $z = j$ was sampled to generate w_i in a document; $P(w_i | z_i = j)$ is the probability that the jth topic was sampled for the ith word token for a given document. Given this, the model specifies the following distribution of words within a document, as shown in Figure 1:

Figure 1: Calculation of the distribution of words within a document

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

To simplify notation, the multinomial distribution of words over topic j -- that is $P(w|z_i = j)$ -- becomes $\varphi^{(j)}$ and the multinomial distribution over topics of document d -- $P(z)$ -- becomes $\theta^{(d)}$. Because of this, the parameters φ and θ determine which words are important for which topics and which topics are important for which documents. To simplify the process of inference of parameters φ and θ, the multinomial distributions of these parameters is approximated using a Dirichlet prior (c.f. Resnik and Hardisty, 2010), with Gibbs Sampling used to infer the parameters φ and θ for all word occurrences (tokens) in all documents (Steyvers and Griffiths 2007), shown in Figure 2:

Figure 2:Calculation of the Dirichlet prior

$$Dir(\alpha_1, \ldots, \alpha_T) = \frac{\Gamma(\Sigma_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

In practice, the distributions θ and φ are matrices representing probabilities of topics by document and tokens by topic respectively (often referred to as mixture weights and mixture components). Setting all the Dirichlet hyperparameters to a fixed value for each distribution (α

for θ and β for φ) results in a smoothing function on topic and word distributions respectively: the choices of the hyperparameters α and β directly influence the discriminative power of the model such that large values of α will result in a model that assumes uniform topic probabilities for each document while low values of α assume more skewed topic distributions for each document. That is, each document is made up of only a few topics. The hyperparameter β performs a similar function for determining the distribution of words against topics (Blei, Ng, and Jordan 2003; Steyvers and Griffiths 2007).

Insight Workflow

Our study was exploratory in approach, interrogating the data to answer broad questions. Because the data had not been previously analysed, we adopted a combination of computational and human-based approaches in an iterative fashion, using our computational tools to generate analyses and also facilitate close reading (Mimno, 2012). This way, the computational and human approaches are able to inform and build upon each other in order to create a greater understanding of the new dataset. In order to achieve insight from the case studies, we therefore needed a consistent workflow that allowed us to develop the natural language processing analyses with strong researcher supervision and interaction. Based on the experience in Mimno (2012) and Mehl and Gill (2010), the steps of the 'insight workflow' employed in the current analysis are shown in Table 1.

Table 1: Stages of our exploratory analysis ('insight workflow')

Stage	**Description**
1. Manual exploration	All team members read a sample of documents to get a good sense of their structure and content.
2. Automatically assisted exploration	Computational techniques used to explore the texts, e.g., extract the most frequent words, grammatical categories and entities, for example people, places, organisations, etc.
3. Hypotheses-driven exploration	Hypotheses based on steps (1) and (2) developed into words, phrases and regular expressions to be tested using KWIC and document search.
4. Topic modelling	LDA was used to identify topics within the documents (method is described in more detail in Table 2).
5. Iteration	Manual and automatically assisted exploration steps were iterated for the interpretation and validation of the topic model (Quinn et al. 2010).

At the centre of our computational analysis was the topic modelling workflow, which included the steps outlined in Table 2.

Table 2: Steps of the topic modelling process

Stage	Description
1. Pre-processing the data	To ensure suitability for input to the topic modelling algorithm.
2. Exploring the data	Generate a number of topic models with different parameters, e.g., number of topics, specified.
3. Examine and validate the resulting 'topic keys' of the different models	Ideally models should demonstrate (a) internal consistency but minimise (b) the repetition/duplication of similar topics or the possible framing of larger topics (Druckmann, 2001) and (c) should not contain large numbers of 'junk' words (if not, it may point to the data not being cleaned or pre-processed appropriately before analysis?; see Boyd-Graber et al 2014 for more detailed discussion). At this stage, human reading and assessment of the topics is critical (AlSumait et al. 2009).
4. Diagnostics	Apply and examine diagnostic measures (such as those generated by Mallet software) to the topics.
5. Iteration	Iterate through steps (2), (3), and (4) to narrow down the number of topics until a satisfactory topic model solution is achieved.
6. Name/interpret topics	An iterative process using manual reading of topic keys and KWIC of topic keys in the original data ensures accurate interpretation of topics.
7. Future analysis using the	Deriving a set of topics that summarises a set of documents

topics (optional)	might be an end in itself for some researchers; however, for others this may simply be a stage that enables future analysis: For example, the difference in use of topics across the dataset might be used to answer a research question, applying the topics derived from the current dataset (specified during generation) to a new dataset.

Results and Discussion

In this section, we describe the application of our insight methodology to the HEFCE REF data as a worked example, illustrating this through the presentation of selected results and discussing the strengths and weaknesses of this approach. Following the first steps of exploring the data through manual reading (IW1 [Footnote 3 here]), we also used standard automated methods to get a sense of the dataset (IW2). For example, a very basic frequency analysis of the most common words in Section 4 (excluding function or stop words [Footnote 4 here]) can be found in Table 3, with Table 4 showing the most frequent words, manually reviewed to be related to policy (note word frequencies vary between analyses due to different tokenisation processes used by the respective software; Z-scores indicate the number of standard deviations from the mean frequency). Both of these demonstrate that such simple information can begin to usefully give a quick overview of key themes involved in certain areas of research.

Table 3: Most frequent words

Word	Frequency	Z-score
research	29,752	13.62
impact	14,449	6.60
work	10,359	4.72
UK	9,742	4.44
new	9,317	4.24
policy	9,223	4.20
development	8,099	3.68
project	7,443	3.38
international	6,734	3.06
public	6,409	2.91

Table 4: Selected keywords related to policy

Word	Frequency	Z-score
policy	9,223	4.20
government	5,251	2.38
parliament	834	0.35
policymakers	375	0.14
house_of_Lords	287	0.10

Extraction of the most frequent countries mentioned in the 'Details of the impact' section is shown in Table 5; providing the basis for subsequent deeper analysis by manual reading (note that in this table and our analysis we omit references to the United Kingdom/UK since our analysis was focused on identifying international collaborations and impacts).

Table 5: Ten most frequent countries mentioned in 'Details of the impact' (Section 4)

Country	**Number of mentions**	**% of total (excluding UK)**
United States	1,822	10
Australia	1,076	6
Canada	878	5
Germany	864	5
France	678	4
Ireland	624	3
China	619	3
Netherlands	603	3
India	492	3
Italy	484	3

Whilst raw word frequencies can indicate the 'popularity' of that word within the corpus, it is also useful to examine how distinctive is that word usage. One way to do this is to compare the usage within the corpus under investigation to a benchmark – or reference – corpus. In Table 6 we demonstrate this, by showing the top ten words used proportionately more ('over used') in

the impact case studies compared against the sample of the BNC spoken corpus for words with frequencies greater than 5 (this analysis was performed using the Wmatrix corpus comparison tool; Rayson, 2008). In the table, O1 refers to observations for the impact case study corpus (raw frequencies, and proportions given the size of the corpus), with O2 referring to the BNC reference corpus, and G2 giving the log likelihood score (which can be interpreted for significance using a chi square table of critical values; '+' in column 6 indicate greater usage in the O1 corpus).

Table 6: Corpus comparison of impact case studies vs BNC sampler written corpus

Word	**O1**		**O2**			**G2**
	Freq.	**Prop.**	**Freq.**	**Prop.**		
research	30266	0.73	186	0.02	+	11046.76
impact	14480	0.35	126	0.01	+	5044.43
has	32623	0.79	2695	0.28	+	3610.43
UK	9747	0.23	207	0.02	+	2764.50
and	165897	4	28384	2.93	+	2507.94
project	7555	0.18	107	0.01	+	2398.64
policy	9223	0.22	263	0.03	+	2339.72
in	110132	2.65	18086	1.87	+	2082.59
public	8474	0.2	279	0.03	+	2011.56
development	8100	0.2	274	0.03	+	1896.92

Here we can see in the impact case study corpus that there are some words which were also identified by the raw frequency analysis (policy, uk, project). However, what is potentially more interesting is the distinctive use of words which would normally be ignored as function words or

stop words in frequency analysis (‘has’, ‘in’), which may indicate the argumentative framework used in many of these case studies, namely: X *has* an impact *in* Y.

In other analysis (IW3), key words, which were hypothesised to be interesting in the context of impact, were searched for in the data. In addition to generating raw frequencies, these could be studied in their surrounding contexts (using the key word in context method) in order to view them (to enhance interpretability, this is most easily conducted on versions of the documents which are not processed for e.g., removal of punctuation, or stop words). An example of this for the keyword stem ‘health’ (which also captures ‘healthy’, ‘healthcare’, etc.) is shown in Table 7.

As can be seen for this example of ‘health’, KWIC shows there are many different possible perspectives that can relate to impact: this is useful in exploration of the data, but is also useful to apply this technique in relation to the topic modelling output in order to explore the meaning and sense of the topics. For example, ‘health’ occurs as a topic key for 6 of the topics in Table 7 (topic numbers 18, 22, 23, 25, 35, 47). In Table 7, it is possible to use KWIC analysis to identify different usage of the term ‘health’ in the context of the surrounding text: lines 1-3 belong to a text using the ‘Food and nutrition’ topic (no. 22, with a proportion of 0.57), whereas lines 4-10 are found in a text which containing the ‘Health care services’ topic (no. 23, proportion of 0.24). As might be expected from the ‘Food and nutrition’ topic, the context also contains the words ‘food’; in the case of the second text (lines 4-10), we do not see other topic keys from ‘Health care services’), but it is clear that is represents a different concern of ‘health’, specifically ‘eye-health’, and also including terms such as ‘inequality’, and ‘commission’. In this brief example we have demonstrated how we would go back to the original text to understand some of the

senses of topics in relationship to words used, and would typically be used for understanding the meanings of topics and the ways in which the topic keys are used. Examining results in light of the original text is a central aspect of the iterative nature of our insight workflow.

Table 7: Key word in context for 'health'

Preceding text	**Keyword**	**Following text**
food, water and energy security. Impacts on	health	and welfare: Public awareness of the merits
the merits of healthy food and the	health	benefits of consuming oats has increased ov
. This coincides with the publication of the	health	claim on oats by the EU [5.3] that
engagement which will lead to improved eye-	health	. B. Increasing awareness through commission
have led directly to community-based Eye	Health	Engagement Projects: (a) RNIB-commissioned
ntervention strategies to address inequalities in	health	care [5]. Using our research it highlights
strategies to reduce this inequality. (b) Public	Health	Action Support Team (PHAST) Eye Care in
strategies employed to reduce inequalities in eye-	health	in ethnic minorities, and proposes remedial

to the development of community-based Eye	Health	Engagement Projects in Glasgow [7] (2012) a
Development Fund (IESD) from the Department of	Health	. The Eye Health Engagement Projects employ

We now focus in more detail on the core of our exploratory insight workflow, topic modelling (IW4): Using the processed data (stemmed, lowercased, and with numbers, punctuation, stop words and non-ascii characters removed, as described in the Method section; corresponding to TM1), the LDA topic modelling package in Mallet (http://mallet.cs.umass.edu) was used to generate models for a variety of number of topics ranging from 10-100 at coarse intervals (e.g. 10, 25, 50, 75, 100); this was to explore the data and to get a sense in human terms of the themes present in the data (TM2) in order to identify a reasonable number of topics to extract. In all topic modelling described here, default parameter settings were used except in the case of α where a relatively low value (0.01) was specified in order to generate topics which relate more distinctly to particular documents (cf. Mimno et al. 2014). Three researchers who were familiar with the data set visually inspected the model outputs to evaluate the 'topic keys' to determine whether they contained many poor topics; specifically ones which were too general, too specific, repetition with other topics, or internally inconsistent (TM3). Topic keys are the top N words generated by the topic modelling algorithm which relate to each topic. These key words can be used to manually interpret the 'theme' or 'themes' described by the topic (e.g., by using KWIC, described above); the number of keywords presented is determined by the researcher, in this case we specified nine which was an artefact of out analysis processing pipeline.

We now iterate through this process (TM5): Once the range of number of topics potentially suitable for the data had been narrowed in range (in this case, between 50 and 100), another set of topic models was generated at finer (10 topic) intervals (e.g. 50, 60, 70, 80, 90, 100; TM2 repeated). These models were then evaluated by hand based on their topic keys and also using diagnostic information generated by Mallet (specified when running each topic model; TM3 and TM4 applied to the second iteration of topic modelling). The diagnostic information was averaged across the topics to give values for each model, and these subsequently plotted alongside each of the other models to provide a broad comparison and to note general trends (e.g., do models appear to perform better or worse with greater or fewer topics in our range). The main diagnostic measures, which were considered to inform our choice of topic model, were: (i) coherence (semantic similarity of words within a topic; greater coherence is better); (ii) distance from corpus (how distinct the topic is relative to the rest of corpus; greater distinctiveness is better); (iii) documents at rank 1 (how many documents this topic best describes; a higher number is better); (iv) distance from uniformity (degree to which probability of a topic relates to a smaller number of words; here lower is better, since we want topics that are most representative of the wider dataset; for example, we would want a topic relating to ‘cars’ to relate to a wider number of words, such as ‘tyre’, ‘gearbox’, ‘engine’, ‘body’, ‘clutch’, ‘Ford’, [and many more], rather than just ‘gearbox’ and ‘Ford’). (For discussion of these measures, see Mimno et al. 2014, and for identifying ‘junk’ or ‘insignificant’ topics – especially those occurring through artefacts of the data format or clean-up process – see AlSumait et al. 2009).

Finally (TM5, iteration 2), the range of topics potentially suitable for our data was narrowed once more (65-80; iteration 2, TM2), and another set of topic models were generated at finer (5

topic) intervals (e.g. 65, 70, 75, 80). Once again, these models were then viewed in relation to their diagnostics and topic keys before finally settling on a single model to describe the data (65 topics; iteration 2, TM3 and TM4).

For 5 of the 65 topics identified we did not assign a topic label; these did not relate to impact of research but rather to generic academic descriptions or vague positive terms relating to the reports ('professor', 'work', 'project', 'improv'), did not indicate a clear topic relating to impact (i.e. could be regarded as insignificant topics; AlSumait et al. 2009), or were incoherent [Footnote 5 here]. We thus ended up with 60 topics for description in the final analysis, shown with their topic keys in the Appendix. In HEFCE (2015), we present these topics in more detail, in particular how different topics relate to the various units of assessments (disciplinary area). To aid interpretation of the topics, each topic is given a human-assigned name, which effectively summarises the meaning of the topic. There is not a definitive process for this step, but we found this was best achieved through human reading of the topic keys (preferably by two or more domain experts), combined with exploring the use of the words represented by the topic keys within the case studies (to ensure correct interpretation of the keys and their respective word senses); in particular, we found that key word in context (KWIC) analysis was especially useful for this process (TM6). It is useful to note that the topic keys are presented in descending order of importance (weighting) for that particular topic, with earlier words/keys being more representative of that topic's meaning. For example, the topic labelled 'Animal husbandry and welfare' (topic number 1) is more relevant to 'anim' or 'welfar' than 'control' or 'uk'. We also note at this stage of interpreting the topics that stemming (i.e. reducing words down to their base form) can introduce some ambiguity to this step: for example, the stem 'commun' relates to topics 12, 36, 49, 52, and 59, and can mean 'community', communities' or 'communication',

'communications'. As discussed above and by examining the other topic keys, this instance of ambiguity was straight forward to resolve (the case in topic 36 solely relates to 'communication'); KWIC in particular has proven to be very valuable for understanding the meaning of the stemmed topic keys.

In terms of summarising the quality of the topic model and its constituent topics, we are confident that this provides a good representation of the data, not only in terms of diagnostic measures, but also equally – and arguably more importantly – in terms of face validity. In addition to the internal consistency of the topics themselves, the very high frequency and widely used topics show frequent words that occur widely across the data (e.g. describing impact and reporting of it), while the others represent the different areas of research and communication. The topics which occur with the greatest frequency in our corpus are shown in Table 8. We discuss the usage of topics in our corpus in more detail in the next section.

Table 8: Top ten topics ranked by proportion in the impact case study corpus

Topic label	**Topic Number**	**Words related to this topic**	**Topic-proportion**
Informing government policy	26	*develop polici nation plan govern inform work strategi assess*	33%
Parliamentary scrutiny	43	*polici govern report public uk committe debat evid commiss*	23%
Community and	12	*local commun project citi council*	19%

local government		*social peopl fund develop*	
Public engagement	46	*peopl particip wai experi comment engag cultur discuss life*	19%
Technology commercialisation	56	*technolog compani develop product univers commerci system market industri*	18%
Media	33	*public bbc media radio programm interview time broadcast articl*	17%
Print media and publishing	45	*univers book intern translat world publish de public uk*	14%
Business and industry	6	*compani busi manag industri product market servic improv sector*	13%
Schools and education	51	*educ school teacher student teach learn univers develop curriculum*	10%
Software development	53	*softwar develop tool system user data model project comput*	9%

Description of Topic Modelling Output [Footnote 6 here]

Although our 65 topic models provided the best solution by human evaluation and diagnostic measures of the various topic models, how well does this topic model make sense of the data? Formal evaluation of topic models is regarded as an aspect of the process, which is not clearly defined (with the exception of junk topics). Therefore in the current study, we evaluated the

suitability and effectiveness of our topic model in relation to objective relevance to our analysis: are there topics which are common across and relevant to a wide range of disciplines; is there face validity in the way some topics relate more or less to particular disciplines; do these topics help us to learn new and unexpected things about the data which could not have been achieved by manual analysis alone? This relates to the seventh step of topic modelling (TM7).

Topic modelling is clearly useful for providing semantic links between the impact case studies based on common concepts, as shown by the topics in the Appendix (Table 9; with count and proportion of topic usage greater than 5% in documents within the corpus shown in columns 4 and 5, respectively), with the top ten topics by overall proportion summarised in Table 8. Of the topics generated by our topic model, we could define a number of what we term 'super topics' that are very common across our data set, and display the breadth and wide relevance of the topics generated by our model. 'Informing government policy' and 'parliamentary scrutiny', for instance, are used across 36% and 27% of the case studies respectively with a proportion greater than five per cent (see column 5 of Table 9). These two topics in particular were also the ones most widespread across all units of assessment; suggesting that researchers from many fields of research contribute to these types of impact. Other super topics relate to impact that is spread across the disciplines, but which still pick up nuances of the different foci present within the different disciplines. For example, 'Technology commercialisation' (identified by the topic keys 'technolog compani develop product univers commerci system market industri') contains case studies from Panels A (Life sciences) and B (Engineering and physical sciences), whilst 'Informing government strategy' occurs in all four Panels, albeit less so in Panel D (Arts and humanities). 'Schools and education' is also distributed across the 4 Main Panels and 36 Units

of Assessment, particularly in Panels C (Social sciences) and D (Arts and humanities). Since topic modelling is a technique, which generates topics based on the whole data set - regardless of the original discipline or assessment panel of the document - this implies that patterns can also be identified across all Units of Assessments, potentially identifying links that a human reader might miss. The above-mentioned three topics are linked to engagement with a general public in economy and society but also relate to most day-to-day activities in academia and thus bring together all disciplines. Commercialisation and policy are often stated outcomes of research projects while teaching activities in universities take a lead in developing education as a whole across disciplines and Units of Assessment.

To consider the spread and specificity of topics across the disciplines, we focus in particular at the Arts and humanities panel impact assessment: Considering the diverse disciplines from philosophy to empirical archaeology and dance incorporated in this panel, it is no surprise that the spread of impact topics is here especially strong: all but four topics -- Laboratory diagnostics, Cancer, Animal husbandry and welfare, and Instrumentation -- occur in at least one of the Panel D units of assessment. Within Panel D, the topic 'Media' is the most commonly occurring one (424 times in the 1627 Arts and humanities panel case studies). The 'Media' topic is rather unspecific: the top nine stemmed topic keys relating to this topic were: 'public bbc media radio programm interview time broadcast articl'. These can incorporate any kind of media impact from appearances on the radio to articles in newspapers. The guidance of the REF before submission was that the impact case studies needed to evidence how they effected change outside academia. Thus, this topic is linked to the attempt to reach a broadly defined public rather than specific bodies. One way to explore in more detail the underlying sense of the Media topic is to identify

other frequently co-occurring topics. For example, the topic often appeared in combination with other more specific topics such as the kind of media used or the topics of the media coverage: 'Literature', Print media and publishing', 'Religion', or 'Public engagement'; this approach can therefore be used to better understand what specific aspect of a broad topic such as Media is referenced in the data, whether in general or in specific case studies.

Finally, we note that our insight workflow combines topic modelling with other analytic techniques, in order to build upon the strength of computational and human methods (IW5). While we note that co-occurrence analysis of topics (described above) can be performed computationally, other analysis requires greater human knowledge and input: for example, in relation to the 'Clinical guidance' topic, external knowledge about the UK healthcare system was used to disambiguate and further explore how impact was achieved within this topic. In particular, knowledge of specific terms such as 'NICE' (the UK National Institute for Health and Care Excellence) and 'QALY' ('Quality Adjusted Life Years' an estimate of health gain from a treatment) were used to form the seeds of keyword search and keyword in context techniques, in order to provide a deeper understanding than that afforded by a particular topic (see HEFCE 2015 for further information of these findings). Here, computational techniques provide support for human analysis. Another example, which required more sophisticated computational techniques, and greater iteration was the process of identifying stakeholders, in particular groups that could be considered beneficiaries or users of the impact described in the 'details of impact' (Section 4) of the case studies. Here in an exploratory process (implemented using NLTK), we extracted nouns, which occurred within the context (here defined as a sentence) following a list of predefined keywords such as 'stakeholders', 'beneficiaries' and/or 'users'. Using this method, a number of nouns could be identified that could be considered beneficiaries. We then conducted

KWIC searching on a selection of the noun groups identified in this way to examine their frequency of occurrences across the different panels. Further reading was, however, needed to disambiguate whether the extracted nouns were part of the impact or just part of the general descriptions in the case studies. One example of what this type of analysis process revealed is that while there are groups potentially benefitting from the case studies relating to their particular field of research - writers benefitting from studies in Panel D, engineers benefitting from studies in Panel B - it is part of the interdisciplinary impact that writers are also mentioned in Panel B and engineers in Panel D.

Therefore as well producing topics which are widely relevant across disciplines in the case of the 'super topics', we have also discovered some which show disciplinary specificity (e.g., to Life sciences rather than Arts and humanities) and that the co-occurrence of topics can uncover new patterns, such as aspects of media engagement. In addition, we have demonstrated how domain expertise can help to dig deeper through the results of the topics back into the data, such as in the case of specific terms (e.g., NICE, QALY), and in identifying patterns of stakeholders across the disciplinary panels.

The highly iterative and interdependent nature of this exercise shows that when exploring complex natural language, it is not the case that computational techniques simply output the final results, but that they assist an iterative process between human analysis and computational exploration and analysis of the data; in addition, this process depended greatly upon the knowledge and expertise of domain experts, in this case in the area of research assessment. As such, this is not simply a case of 'turning the handle', but more a case of more efficient

harnessing of human time and expertise to better analyse a wider range and scale of data. In cases where the same analytic procedures are repeated over large amounts of data, then we expect that greater automation of the whole process would be possible. The strength of this approach is also its weakness: the high level of interactivity between researchers produces richer insights benefitting from the various expertise of the individuals, but demands a grouping of individuals with both technical skills and domain knowledge as well a close working practices; a fundamental assumption is that there are large quantities of relevant data in an appropriate format. In future work, we would be keen to apply our approach to a wider variety of datasets, in particular comparing our approach with domain specific coding frameworks; in addition, it would be interesting to test the applicability of the topics derived from the HEFCE impact case studies to other descriptions of research application and impact.

Conclusions

We have demonstrated the underlying methods and challenges behind exploring and analysing a real-life dataset using a combination of computational and human approaches. In particular, we needed to develop an 'insight workflow' that allowed us to dig deeply and also broadly into the focus of this study, which is 'impact'. The impact case studies needed to be closely analysed with the help of human evaluators, where the challenge was to design an interaction between computing labour and human analysis that would be most effective for our work. We developed the insight workflow based on previous experience, where manual exploration followed automatically assisted methods, which was then deepened through hypothesis-driven exploration and topic modelling. Notably, the computational analysis – especially from topic modelling - provided an empirical entry-point for the more in-depth and labour intensive human analysis of

the texts. By using our combined approach, human effort could most efficiently be directed to the task.

Our approach was able to identify topics relating to impact that are spread across various REF panels and disciplines; this indicates its greater suitability to our varied dataset, compared to traditional methods based on domain-specific predefined classification systems. The topic modelling work in particular was able to find new connections and also developed our conceptual understanding. In particular, we were able to identify semantic links across case studies based on common concepts, to demonstrate the spread of impact topics across and within the disciplines of research, show the existence of impacts not predictable from a particular discipline, and finally, to find specific types of information such as beneficiary groups from the impact of research.

To summarise, we have combined the strengths of computational approaches from natural language processing and text mining with those of human analysis in order to demonstrate how they can explore a dataset: in this case impact case study documents. This approach has given us new insights into the reach of research impact beyond academia, which would have been difficult to obtain using individual methods in isolation. In the detailed description of our insight workflow and methods, we provide a basis for future research by enabling others to build upon our approach in this and other domains.

Footnotes

(1) We note that in topic modelling, the term 'topic' has a particular meaning, indicating one of a number of 'themes' that are automatically extracted from a set of documents, based on word co-occurrence within the documents. We describe the process of topic modelling in more detail in the section 'Computational Techniques used in the Analysis'.

(2) TM1 refers to the Topic Modelling step 1, which relates to the step in our analysis corresponding with those described in Table 2.

(3) IW1 refers to Insight Workflow step 1, indicating the part of the analysis process which relates to the steps described in Table 1.

(4) When stop words were not excluded from frequency analysis, the top ten most frequent words were: 'the', 'of', 'and', 'in', 'to', 'a', 'for', 'on', 'by', 's'.

(5) These five excluded topics were: 1) 'result year benefit increas improv provid cost time signific'; 2) 'sourc work refer import report evid section inform professor'; 3) 'work develop practic intern group confer project profession organis'; 4) public includ websit scienc uk onlin year event engag'; 5) 'remov text public product glass skin durham industri manufactur'.

(6) We note that the results presented here have previously been reported in the HEFCE report *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*. Research Report 2015/01. (HEFCE,

2015). Whilst previously the results were reported in relation to describing impact, here we use them to illustrate the usefulness of our methods. We refer the reader to the HEFCE report for a fuller presentation and description of impact in UK HEIs.

References

AlSumait, L., Barbara, D., Gentle, J., and Domenico, C. (2009). Topic Significance Ranking of LDA Generative Models. In Buntine, W., Grobelnik, M., Mladenić, D., and Shawe-Taylor, J. (Eds.). *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I* (pp. 67-82). Springer-Verlag: Berlin, Heidelberg.

Anthony, L. (2016). *AntConc (Version 3.4.4) [Computer Software]*. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Blanke, Tobias. *Digital Asset Ecosystems: Rethinking crowds and cloud.* Amsterdam: Elsevier, 2014.

Blei, David. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities 2(1),* 8-11.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In Cohen, W. and Moore, A. (Eds.). *Proceedings of the 23rd international conference on Machine learning*. (ICML '06) (pp. 113–120). ACM, New York, NY.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borgman, C.L. (2009). The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly,* Volume 3 Number 4. Available from: http://digitalhumanities.org/dhq/vol/3/4/000077/000077.html

Jordan Boyd-Graber, David Mimno, and David Newman. (2014). Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In Airoldi, E.M.,Blei, D., Erosheva,

E.A., and Fienberg, S.E. (Eds). *Handbook of Mixed Membership Models and Their Applications.* Boca Raton, Florida : CRC Press.

Brauer, R. and Dymitrow, M. (2013). *Using topic modelling to analyse EU's Rural Development policy.* Paper presented at Systematizing and digitalizing Nordic policy studies: Emergent perspectives within Swedish and Finnish research Symposia, Aalto University, 27 November 2013, Helsinki, Finland.

Chaney, A. J.-B., and Blei, D. M. (2012). Visualizing topic models. In Breslin, J.G., Ellison, N.B., Shanahan, J.G., and Tufekci, Z. (Eds.). *Proceedings of the 6th International. AAAI Conference on Weblogs and Social Media* (pp. 419–422). Palo Alto, California: The AAAI Press.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C. and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. In Bengio, Y., Schuurmans, D., Lafferty, J. D.,Williams, C. K. I., and Culotta, A. (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288-296). Red Hook, NY: Curran Associates, Inc.

Druckman, J. N. (2001). The implications of framing effects for citizen competence. *Political Behavior*, *23*, 3, 225–256.

Graham, S., Milligan, I., and Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope, a co-written manuscript by Shawn Graham, Ian Milligan, and Scott Weingart.* Available at: www.http://www.themacroscope.org [last accessed 3 June 2015].

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42*(1):7-15

Harvey, M., Crestani, F., and Carman, M.J. (2013). Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (CIKM '13) (pp. 2309-2314). New York, NY: ACM. DOI=http://dx.doi.org/10.1145/2505515.2505642

HEFCE (March 2015). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies. Research Report 2015/01*. Retrieved from: http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/Analysis,of,REF,impact/Analysis_of_REF_impact.pdf

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning, 95(3),* 423-469.

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89-106.

Kirschenbaum, M.G. (2007). The Remaking of Reading: Data Mining and Digital Humanities. Paper presented at the National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, 12 October 2007. Retrieved from: http://www.csee.umbc.edu/~hillol/NGDM07/abstracts/talks/MKirschenbaum.pdf [last accessed 27 February 2015].

Klein, L., and Eisenstein, J. (2013). Reading Thomas Jefferson with TopicViz: Towards a Thematic Method for Exploring Large Cultural Archives. *Scholarly and Research Communication 4*(3), Article ID 0301121. Retrieved from http://src-online.ca/index.php/src/article/view/121/259.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 2 (1997), 211–240.

Lowe, W. (2004). Content analysis and its place in the (methodological) scheme of things. *Qualitative Methods*, *2*(1), 25–27.

McDonald, S. (2000). *Environmental Determinants of Lexical Processing Effort*. Unpublished Ph.D. thesis, University of Edinburgh.

Mehl, M. and Gill, A. (2010). Computerized content analysis. In S. Gosling and J. Johnson, (Eds.), *Advanced Methods for Behavioral Research on the Internet*. Washington, DC: American Psychological Association Publications,.

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *ACM Journal of Computing in Cultural Heritage*, 5, 1, Article 3. DOI=http://dx.doi.org/10.1145/2160165.2160168

Moretti, F. (2013). *Distant reading*. London: Verso Books..

Morgan Jones, M., and Grant, J. (2013). Making the Grade. Methodologies for Assessing and Evidencing Research Impact. In Dean, A., Wykes, M., and Stevens, H. (Eds.), *7 Essays on Impact. DESCRIBE Project Report for Jisc* (pp. 25-43). Exeter: University of Exeter Retrieved from http://www.exeter.ac.uk/media/universityofexeter/research/ourresearchexcellence/describeproject/pdfs/2013_06_04_7_Essays_on_Impact_FINAL.pdf

Oberlander, J. and Gill, A.J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42,* 239-270.

Pennebaker, J. W., Mehl, M. R., and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., and Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics. 13*(4), 519-549.

REF (2011). Assessment Framework and Guidance on Submissions. REF 02.2011 (as of 28 July 2014) Available from www.ref.ac.uk/pubs/2011-02/ [last accessed 27 February 2015].

REF (2012). Panel Criteria and Working Methods. REF 01.2012 (as of 28 July 2014). Available at: www.ref.ac.uk/pubs/2012-01/ [last accessed 27 February 2015].

Resnik, P. and Hardisty, E. (2010). *Gibbs Sampling for the Uninitiated*. Technical Report CS-TR-4956, UMIACS-TR-2010-04, LAMP-153, University of Maryland.. Available from https://www.umiacs.umd.edu/~resnik/pubs/LAMP-TR-153.pdf

Rule, A., Cointet, J.-P. and Bearman, P.S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014. *Proceedings of the National Academy of Sciences,, 112*(35),10837-44.

Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*, Laurence Erlbaum, 2007.

Sula, C. A. (2013). Digital Humanities and Libraries: A Conceptual Model. *Journal of Library Administration, 53*, 10–26.

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., Leenders, A. G. M. and McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods, 8*, 443-444.

Vasalou, A, Gill, AJ, Mazanderani, F, Papoutsi, C and Joinson, A (2011). Privacy dictionary: A new resource for the automated content analysis of privacy, *Journal of the American Society for Information Science and Technology, 62(11),* pp. 2095-2105.

Yi-Chia Wang, Moira Burke, and Robert E. Kraut. 2013. Gender, topic, and audience response: an analysis of user-generated content on Facebook. In Mackay, W.E., Brewster, S., and Bødker, S. (Eds.). *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13) (pp. 31-34). New York, NY: ACM.

Appendix

Table 9: Topic model displaying topics and top topic key words (stemmed)

Topic label	Topic Number	Words related to this topic	No. docs >5%	% > 5%
Animal husbandry and welfare	1	*anim welfar farm veterinari breed diseas control uk farmer*	143	2%
Architecture and building	2	*design build construct standard industri structur project architectur engin*	230	3%
Arts and culture	3	*art artist work cultur creativ project public audienc exhibit*	417	6%
Asia	4	*china chines india arab indian asian intern east foreign*	154	2%
Banking, finance and monetary policy	5	*bank financi polici econom financ credit tax risk central*	223	3%
Business and industry	6	*compani busi manag industri product market servic improv sector*	847	13%
Cancer	7	*cancer patient treatment clinic trial uk breast guidelin therapi*	260	4%
Children, young people and families	8	*children child young parent famili imp programm work support*	412	6%
Climate change	9	*climat chang energi carbon emiss uk environment adapt wast*	281	4%
Clinical guidance	10	*guidelin patient clinic treatment recommend stroke nice risk trial*	480	7%

Clinical tests	11	*test patient clinic genet diseas diabet diagnosi diagnost treatment*	321	5%
Community and local government	12	*local commun project citi council social peopl fund develop*	1318	20%
Computing and quantum physics	13	*comput secur light ibm physic intel scienc particl imag*	127	2%
Crime and justice	14	*polic crime prison justic xxxx offic violenc offend victim*	259	4%
Cultural and heritage preservation	15	*heritag archaeolog site visitor histor museum project cultur tourism*	259	4%
Defence and security	16	*defenc militari secur war conflict uk forc arm offic*	156	2%
Democracy and political engagement	17	*polit elect parti democraci elector vote candid poll pd*	112	2%
Dentistry	18	*kcl dental drug oral treatment king prof scott health*	137	2%
Engineering, design and manufacturing	19	*engin design process manufactur fuel develop materi industri improv*	341	5%
Europe	20	*european eu europ intern commiss polici human countri state*	446	7%
Film and theatre	21	*film theatr perform plai audienc product festiv screen director*	242	4%
Food and nutrition	22	*food product industri nutrit health crop agricultur uk seed*	255	4%
Health care services	23	*health care servic nh hospit patient nation improv practic*	787	12%
Historical archives	24	*histori archiv public histor project librari heritag cultur materi*	565	8%
Infectious diseases control	25	*malaria control health diseas resist infect treatment drug programm*	113	2%

Informing government policy	26	*Develop polici nation plan govern inform work strategi assess*	2417	36%
Instrumentation	27	*laser instrument materi product process imag manufactur develop industri*	283	4%
International development	28	*develop countri intern world africa polici global govern african*	542	8%
Laboratory diagnostics	29	*test assai diagnost dna detect protein laboratori sequenc develop*	207	3%
Law and justice	30	*law legal court justic case judg act legisl lawyer*	286	4%
Literature	31	*book read poetri write literari writer publish literatur translat*	267	4%
Marine and ocean science	32	*marin fish fisheri sea coastal ship ocean manag environment*	143	2%
Media	33	*public bbc media radio programm interview time broadcast articl*	1309	20%
Medical ethics	34	*ethic disabl human transplant cell donat donor uk medic*	160	2%
Mental health	35	*mental health clinic servic train treatment intervent patient psycholog*	409	6%
Mobile technologies	36	*mobil system technolog network servic digit app phone commun*	197	3%
Modelling and forecasting	37	*model data method statist forecast predict estim risk measur*	341	5%
Museums and exhibitions	38	*exhibit museum visitor art galleri collect curat displai public*	353	5%
Music, dance and performance	39	*music perform danc work sound audienc concert record festiv*	184	3%
Nature and conservation	40	*conserv natur manag forest land speci biodivers environment project*	229	3%

Nuclear energy	41	*nuclear power energi nois electr system industri monitor oper*	227	3%
Oil and gas	42	*oil ga space explor industri model field bp mission*	191	3%
Parliamentary scrutiny	43	*polici govern report public uk committe debat evid commiss*	1824	27%
Pharmaceuticals	44	*drug develop pharmaceut trial compani clinic phase discoveri industri*	397	6%
Print media and publishing	45	*univers book intern translat world publish de public uk*	732	11%
Public engagement	46	*peopl particip wai experi comment engag cultur discuss life*	1223	18%
Public health and prevention	47	*health screen hiv vaccin women programm recommend prevent nation*	270	4%
Regional innovation and enterprise	48	*innov busi region sme birmingham enterpris support programm univers*	209	3%
Regional languages of British Isles	49	*languag ireland wale welsh northern cardiff irish english commun*	212	3%
Religion	50	*church religi christian religion faith cathol spiritu confer bibl*	147	2%
Schools and education	51	*educ school teacher student teach learn univers develop curriculum*	765	11%
Scotland	52	*scottish scotland glasgow edinburgh govern aberdeen public dunde commun*	192	3%
Software development	53	*softwar develop tool system user data model project comput*	760	11%
Sports	54	*sport game coach footbal athlet olymp perform physic player*	164	2%
Surgery, implants and devices	55	*patient clinic surgeri hospit medic imag implant surgic devic*	224	3%

Technology commercialisation	56	*technolog compani develop product univers commerci system market industri*	1454	22%
Transport	57	*transport safeti road rail risk fire oper train uk*	196	3%
Water and flood management	58	*water flood environ risk manag environment uk qualiti pollut*	234	4%
Women, gender, and minorities	59	*women equal gender migrat divers ethnic commun group refuge*	266	4%
Work, labour and employment	60	*employ union labour trade work worker wage employe social*	159	2%

Table 10: REF2014 Main panels and units of assessment

Main panel	**Unit of assessment**	
A	1	Clinical Medicine
	2	Public Health, Health Services and Primary Care
	3	Allied Health Professions, Dentistry, Nursing and Pharmacy
	4	Psychology, Psychiatry and Neuroscience
	5	Biological Sciences
	6	Agriculture, Veterinary and Food Science
B	7	Earth Systems and Environmental Sciences
	8	Chemistry
	9	Physics
	10	Mathematical Sciences
	11	Computer Science and Informatics
	12	Aeronautical, Mechanical, Chemical and Manufacturing Engineering
	13	Electrical and Electronic Engineering, Metallurgy and Materials
	14	Civil and Construction Engineering
	15	General Engineering
C	16	Architecture, Built Environment and Planning
	17	Geography, Environmental Studies and Archaeology
	18	Economics and Econometrics
	19	Business and Management Studies
	20	Law
	21	Politics and International Studies
	22	Social Work and Social Policy
	23	Sociology
	24	Anthropology and Development Studies
	25	Education
	26	Sport and Exercise Sciences, Leisure and Tourism
D	27	Area Studies
	28	Modern Languages and Linguistics
	29	English Language and Literature
	30	History
	31	Classics
	32	Philosophy
	33	Theology and Religious Studies
	34	Art and Design: History, Practice and Theory
	35	Music, Drama, Dance and Performing Arts
	36	Communication, Cultural and Media Studies, Library and Information Management

Table 11: Samples of Impact Case studies analysed ('Details of the impact' section)

The Durham research described in Section 2 has been transferred to industry through three different business models (income generated growth, corporate venturing, and venture capital). The transfer methods are summarized in the flow chart below and an example of impact generated through each method is given in the following sections. *(a) Income Generated Growth (Surface Innovations Ltd, Durham, UK):* Atomized-Spray Plasma Deposition (ASPD) described in [6] is capable of producing a wide variety of thin, high quality, functional coatings, at throughputs attractive to a large number of markets. The approach allows lost-cost substrates to exhibit the surface properties and performance of far more expensive materials. To exploit this technology Badyal and Dr Luke Ward (a former PhD student) founded the IP-ownership company Surface Innovations Ltd. in 2001 [Im1]. Durham University agreed to assign non-industrially sponsored intellectual property developed within the Badyal group to the company in return for an equity stake. 14 core patent families were filed during the period 2001-2010 on surface functionalization for applications including: filtration; antifogging; bioarrays; antibacterial; antifouling; high dielectric constant; super-repellency; fog harvesting; and rewritability. The company was funded by loans and income generated from prototype development amounting to £824K for industrial partners including: Siemens (Germany); Arcelor (Belgium); Procter & Gamble (USA); (accessed from http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=11778)
The evidence of the research impact is wide and significant. It has been evidenced by a) advancing assistive technologies in a sporting environment; b) having peer-reviewed knowledge; c) being of an applied nature with out-reach to athletes, coaches and wheelchair manufacturers and d) has instigated collaborative networks with several key internationally renowned researchers. The PHC's research (2008-13) has delivered on a number of performance related projects which have impacted on UK Sports funded Paralympic Sports and contributed to ParalympicGB's achievements at the London 2012 Paralympics **[5.1, 5.2]**. Supporting Paralympic Performance: The outputs from the PHC's wheelchair configuration research theme has significantly influenced the preparation strategies of the Paralympic athletes leading into the 2012 London Paralympic Games, by better educating them about wheelchair configuration and chair choice (all members of the wheelchair rugby and basketball teams, n=36). The research findings have been presented by Dr's Goosey-Tolfrey and Mason to ParalympicsGB practitioners and both GB athletes/coaches. The LU based research found that changes in wheelchair camber greatly influenced wheelchair mobility performance. With Dr's Goosey-Tolfrey and Mason's assistance they worked closely with UK Sport's Research and Innovation team on individual case studies to optimise straight line speed and agility within the sport of wheelchair rugby. (accessed from http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=42396)
The LSTM's work on strategies to improve access to TB diagnosis and treatment was driven by our pro-poor and equity perspectives and our understanding of the barriers that prevent

disenfranchised populations accessing healthcare services. These research programmes have directly influenced policy and local TB control programmes practice.

Policy: Achieving policy change in TB healthcare at the international level requires primary research evidence, time and extensive engagement with researchers, policymakers and funding agencies. Squire and Cuevas were invited to the WHO's annual Strategic and Technical Advisory Group (STAG-TB) meetings to discuss new evidence and give perspectives on TB, SM, poverty and access to services in 2009, 2010, 2011 2012 and 2013. Cuevas served as the Chair of the Stop TB Partnership New Diagnostics Working Group on SM (2007-2011). These contributions yielded numerous contributions to policy and practice. Including the STOP TB Departments adoption of a new milestone within its End of TB strategy, `No families should face catastrophic health costs as a result of TB'. As documented in the slides of the 65th World Health Assembly in May 2012, when Member States including Brazil, UK, Italy, Swaziland, Saudi Arabia and others, called upon the WHO to develop a new post-2015 TB strategy.

(accessed from http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=3147)

i1) Literary heritage of a major British author is adapted and interpreted for the reading public, simultaneously contributing to the economic prosperity of the creative industries in a remote region of the UK.

Sellers' *Vanessa and Virginia* was published in June 2008 by Two Ravens Press, a small Scottish independent literary publishing company. Founded in 2006 near Ullapool in the Highlands, Two Ravens has since relocated to Uig on the Isle of Lewis. *Vanessa and Virginia* retails for £8.99 and at the end of assessment period had sold around 2,100 copies and around 50 e-books in the UK, most orders being taken directly through Two Ravens' website. [S1] According to the Director of Two Ravens: `*This makes it our bestselling work of fiction, and Two Ravens Press' bestselling book ever. It is unusual for a small literary press to sell more than 1000 copies of a work of literary fiction, especially by a first-time novelist, and so the fact that Vanessa and Virginia has done so well (and, on the strength of our UK publication, been sold on to the USA and several other countries) has meant that we have been able to use this success story as inspiration to other authors we publish.*

(accessed from http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=35306)